ORIE 4741: Learning with Big Messy Data

Exploratory Data Analysis

Professor Udell
Operations Research and Information Engineering
Cornell

September 13, 2021
Announcements

▶ If you’re taking lecture async: remember to submit participation post after each class!
▶ Otherwise, register your iClicker.
▶ Sections start next week. They are optional, attend any one you prefer. Only two (on Tuesday and Wednesday) will be live.
▶ Office hours: Zoom links or rooms and times are posted on course website.
▶ Gradescope is open for submission of hw0, due Thursday 9:30am.
▶ First quiz this week! It should occupy about 20 minutes; you’ll have up to half an hour to complete it. Start it anytime between 10am Friday and noon Saturday.
Questions from zulip

- enrollment: yes, we expect you’ll get in!
- protocol:
  - use the right stream (eg, general, homework, project, ...) and a good subject line
  - search for your question before posting new question
Our programming language policy

- we’ll do demos and provide homework starter code in python
- you’re welcome to use any language you like (that your TAs can read) for homework or project
- TAs will only support python
Topics to review

We will cover (most of) these in section, too:

▶ Linear algebra: invertible matrices, rank, norm, basic matrix identities. When is a matrix invertible?
▶ QR factorization
▶ Gradients (multivariate derivative)
▶ Projections
▶ SVD
▶ Maximum likelihood estimation
▶ Union bound
▶ Computational complexity
Why look at the data?

- detect errors in data
- check assumptions
- select appropriate models
- understand relationships among the features
- understand relationships between features and labels
How to look at the data?

- inspect raw data
- summary statistics
- visualize
American community survey

2013 ACS:

▶ 3M respondents, 87 economic/demographic survey questions
  ▶ income
  ▶ cost of utilities (water, gas, electric)
  ▶ weeks worked per year
  ▶ hours worked per week
  ▶ home ownership
  ▶ looking for work
  ▶ use foodstamps
  ▶ education level
  ▶ state of residence
  ▶ ...

▶ 1/3 of responses missing

find it at https://people.orie.cornell.edu/mru8/orie4741/data/acs_2013.csv
How do computers work?

on a laptop:

▸ hard disk: usually $\leq 500$ GB
▸ memory (RAM): usually $\leq 16$ GB
▸ many programs (e.g., Excel): substantially more limited

how big is ACS data?

$3,000,000$ respondents $\times 100$ questions = $300,000,000$ numbers $\approx 300$ MB
How do computers work?

on a laptop:

- hard disk: usually $\leq 500$ GB
- memory (RAM): usually $\leq 16$ GB
- many programs (e.g., Excel): substantially more limited

don’t load a giant file into memory.
your computer will crash.
How do computers work?

on a laptop:

▶ hard disk: usually $\leq 500$ GB
▶ memory (RAM): usually $\leq 16$ GB
▶ many programs (e.g., Excel): substantially more limited
don’t load a giant file into memory.
your computer will crash.

how big is ACS data?
3M respondents $\times$ 100 questions $= 300M$ numbers $\approx 300$MB
Inspect raw data

solution for large files: technology from the 70s!

bash shell:

- “how big are these files?”: ls -lh
- “show me some lines from the file”: head, tail, less
- “how many lines are in the file?”: wc -l
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHTYPE</td>
<td>household type</td>
<td>categorical</td>
</tr>
<tr>
<td>STATEICP</td>
<td>state</td>
<td>categorical</td>
</tr>
<tr>
<td>OWNERSHP</td>
<td>own home</td>
<td>Boolean</td>
</tr>
<tr>
<td>COMMUSE</td>
<td>commercial use</td>
<td>Boolean</td>
</tr>
<tr>
<td>ACREHOUS</td>
<td>house on $\geq 10$ acres</td>
<td>Boolean</td>
</tr>
<tr>
<td>HHINCOME</td>
<td>household income</td>
<td>real</td>
</tr>
<tr>
<td>COSTELEC</td>
<td>monthly electricity bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTWATR</td>
<td>monthly water bill</td>
<td>real</td>
</tr>
<tr>
<td>COSTGAS</td>
<td>monthly gas bill</td>
<td>real</td>
</tr>
<tr>
<td>FOODSTMP</td>
<td>on food stamps</td>
<td>Boolean</td>
</tr>
<tr>
<td>HCOVANY</td>
<td>have health insurance</td>
<td>Boolean</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>currently in school</td>
<td>Boolean</td>
</tr>
<tr>
<td>EDUC</td>
<td>highest level of education</td>
<td>ordinal</td>
</tr>
<tr>
<td>GRADEATT</td>
<td>highest grade level attained</td>
<td>ordinal</td>
</tr>
<tr>
<td>EMPSTAT</td>
<td>employment status</td>
<td>categorical</td>
</tr>
<tr>
<td>LABFORCE</td>
<td>in labor force</td>
<td>Boolean</td>
</tr>
<tr>
<td>CLASSWKR</td>
<td>class of worker</td>
<td>Boolean</td>
</tr>
<tr>
<td>WKSWORK2</td>
<td>weeks worked per year</td>
<td>ordinal</td>
</tr>
<tr>
<td>UHRSWORK</td>
<td>usual hours worked per week</td>
<td>real</td>
</tr>
<tr>
<td>LOOKING</td>
<td>looking for work</td>
<td>Boolean</td>
</tr>
<tr>
<td>MIGRATEE1</td>
<td>migration status</td>
<td>categorical</td>
</tr>
</tbody>
</table>
Python is a programming language: it parses human-readable code to machine-readable code, executes it, returns the answer.

Jupyter is a protocol for interacting with a programming language.

Jupyter stores inputs and outputs as .ipynb files.

Jupyter notebooks display inputs and outputs in a browser.

Google Colab is an interface to a webserver running Python.
Python and Jupyter

▸ Python is a programming language: it parses human-readable code to machine-readable code, executes it, returns the answer

▸ Jupyter is a protocol for interacting with a programming language.

▸ Jupyter stores inputs and outputs as .ipynb files.

▸ Jupyter notebooks display inputs and outputs in a browser

▸ Google Colab is an interface to a webserver running Python

how to access?

▸ install Python with Anaconda distribution (versions 3.7 or 3.8 are fine)

▸ use Google Colab
Summary statistics

univariate

▶ mean, median, mode
▶ max, min, range
▶ variance
▶ ...

explore via Python + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb
Summary statistics

univariate

- mean, median, mode
- max, min, range
- variance
- ...

explore via Python + Jupyter notebook

https://github.com/ORIE4741/demos/blob/master/eda.ipynb

multi- (but usually just bi-) variate

- correlation, covariance
- ...

The perils of summary statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
<td></td>
</tr>
</tbody>
</table>
The perils of summary statistics

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>

same mean, variance, correlation, line of best fit...
The perils of summary statistics

- $Y_1$ vs $X_1$
- $Y_2$ vs $X_2$
- $Y_3$ vs $X_3$
- $Y_4$ vs $X_4$
The perils of summary statistics: modern update

https://www.autodeskresearch.com/publications/samestats
What to visualize?

- examples across all features (usually not)
- plot features across all examples (much more common)
Best practices

- Always label your axes.
- Ensure all marks on plot are meaningful.
- Beware of pie charts; bar charts are often easier to read.
- Beware of line plots; if your data is not continuous, try scatter plot instead.
- Consider the scale of your axes. Log scale or not?
- Consider which curves to plot on same axes. Make comparisons easy!
**Beware of bad data**

<table>
<thead>
<tr>
<th>Value</th>
<th>Value Label</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Weighted Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 30</td>
<td>Number of days</td>
<td>159,327</td>
<td>36.43</td>
<td>35.59</td>
</tr>
<tr>
<td>88</td>
<td>None</td>
<td>269,145</td>
<td>61.53</td>
<td>62.53</td>
</tr>
<tr>
<td>77</td>
<td>Don’t know/Not sure</td>
<td>7,602</td>
<td>1.74</td>
<td>1.58</td>
</tr>
<tr>
<td>99</td>
<td>Refused</td>
<td>1,336</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>BLANK</td>
<td>Not asked or Missing</td>
<td>26</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Take away

- always look at (some of) your data
- decide what question you want to answer
Questions?

https://docs.google.com/spreadsheets/d/1vLbwi0WCOon0wU6cU_r0RHAoY7C0fDZ1F8Yq09pqYYuk